

# Preliminary Evidence on the Effectiveness of Psychological Treatments Delivered at a University Counseling Center

Takuya Minami  
University of Utah and Northeastern University

D. Robert Davies  
University of Utah

Sandra Callen Tierney  
University of Wisconsin Medical School

Joanna E. Bettmann, Scott M. McAward,  
Lynnette A. Averill, Lois A. Huebner,  
and Lauren M. Weitzman  
University of Utah

Amy R. Benbrook  
Northeastern University

Ronald C. Serlin and Bruce E. Wampold  
University of Wisconsin—Madison

Treatment data from a university counseling center (UCC) that utilized the Outcome Questionnaire–45.2 (OQ-45; M. J. Lambert et al., 2004), a self-report general clinical symptom measure, was compared against treatment efficacy benchmarks from clinical trials of adult major depression that utilized similar measures. Statistical analyses suggested that the treatment effect size estimate obtained at this counseling center with clients whose level of psychological distress was above the OQ-45 clinical cutoff score was similar to treatment efficacy observed in clinical trials. Analyses on OQ-45 items suggested that clients elevated on 3 items indicating problematic substance use resulted in poorer treatment outcomes. In addition, clients who reported their relational status as separated or divorced had poorer outcomes than did those who reported being partnered or married, and clients reporting intimacy issues resulted in greater numbers of sessions. Although differential treatment effect due to training level was found where interns and other trainees had better pre–post outcome than did staff, interpretation of this result requires great caution because clients perceived to have complicated issues are actively reassigned to staff. More effectiveness investigations at UCCs are warranted.

*Keywords:* university counseling centers, benchmarking, treatment effectiveness, therapist effect, clinical assessment

The front-line mental health services for many students pursuing higher education are their college and university counseling centers (UCCs). Although UCCs fulfill many roles—including training, workshops, outreach presentations, and consultation—more time is spent on providing direct counseling services to clients than on any other single activity. According to the 2008 National Survey of Counseling Center Directors (Gallagher, 2009), which included the participation of 284 UCCs mainly from the United States, counselors were expected to spend an average of 61.8% of

their time providing direct service to students, with an average of 23.0 client hours per week. Considering that an average of 9.0% of enrolled students seek counseling in a year, it is evident that UCCs are one of the major providers of mental health services.

The continuous flow of students seeking counseling at these academic institutions has naturally drawn numerous clinical studies to UCCs over the years. The range of issues being studied has been broad, including counseling processes (e.g., Davies, Burlingame, Johnson, Gleave, & Barlow, 2008; Kahn, Achter, & Sham-

---

Takuya Minami, Department of Educational Psychology and University Counseling Center, University of Utah, and Department of Counseling and Applied Educational Psychology, Northeastern University; D. Robert Davies, Scott M. McAward, Lois A. Huebner, and Lauren M. Weitzman, University Counseling Center, University of Utah; Sandra Callen Tierney, Wisconsin Psychiatric Institute and Clinics, Department of Psychiatry, University of Wisconsin Medical School; Joanna E. Bettmann, College of Social Work, University of Utah; Lynnette A. Averill, Department of Educational Psychology, University of Utah; Amy R. Benbrook, Department of Counseling and Applied Educational Psychology, Northeastern University; Ronald C. Serlin, Department of Educational Psychology, University of Wisconsin—Madison; Bruce

E. Wampold, Department of Counseling Psychology, University of Wisconsin—Madison.

An earlier version of this article was presented at the 113th Annual Convention of the American Psychological Association, Washington, DC. We are greatly indebted to all the clients, therapists, and administrative staff members who have supported the University Counseling Center at the University of Utah throughout the years. We sincerely hope that our results will lead to further benefits to our clients at the center as well as clients at other university counseling centers.

Correspondence concerning this article should be addressed to Takuya Minami, Department of Counseling and Applied Educational Psychology, Northeastern University, 209 Lake Hall, 360 Huntington Avenue, Boston, MA 02115-5000. E-mail: t.minami@neu.edu

baugh, 2001; Kivlighan, McGovern, & Corazzini, 1984; Tracey, Sherry, & Albright, 1999), dose-effect relationships (e.g., Draper, Jennings, Baron, Erdur, & Shankar, 2002; Erdur, Rude, & Baron, 2003; Wolgast et al., 2005), feedback (e.g., Lambert et al., 2001), psychometric evaluation (e.g., Hayes, 1997), relationship between satisfaction and outcome (e.g., Tracey, 1989), social support (Mallinckrodt, 1989), therapist effects (e.g., Okiishi, Lambert, Eggett, Nielsen, Dayton, & Vermeersch, 2006; Okiishi, Lambert, Nielsen, & Ogles, 2003), and transportability of empirically supported treatments into UCCs (e.g., Hogg & Deffenbacher, 1988). As such, in addition to providing clinical services, UCCs have contributed to advancing knowledge in psychotherapy through research.

However, despite the breadth of issues investigated at UCCs, it is rather surprising that only a handful of empirical studies specifically have targeted their investigation on the overall effectiveness of treatments provided in this setting (e.g., Snell, Mallinckrodt, Hill, & Lambert, 2001; Vonk & Thyer, 1999; Wilson, Mason, & Ewing, 1997). In other words, most investigations that have taken place at UCCs have been just that—studies that utilized UCCs as the research setting rather than the focus of the study. For example, Tracey et al. (1999) investigated the relationship between counselor–client behavior complementarity and clinical outcomes at a UCC. This study was meticulously conducted, attending to the concern for quality of treatment by including observer session ratings of the cognitive therapy provided. However, because Tracey et al. were not concerned with the absolute level of treatment effectiveness, no comparison of the treatment effects with external criteria were reported. Similarly, Kahn et al. (2001) studied the effect of client distress disclosure on treatment outcome. Although Kahn et al.'s pre–post *t* tests on perceived stress and symptomatology were statistically significant, again no context was provided to evaluate the absolute magnitude of these effects. Thus, because most studies conducted at UCCs have focused on research questions other than pre- and posttreatment outcome, these studies generally do not provide an indication of how effective the treatment has been compared with that from other forms of mental health services.

The few studies that do directly focus on treatment effectiveness at UCCs provide some indications of effectiveness (e.g., Snell et al., 2001; Vonk & Thyer, 1999; Wilson et al., 1997). However, because these studies vary significantly in their methodology, they are difficult to synthesize. For example, using student retention as an indicator of effectiveness, Wilson et al. (1997) followed up on all students who requested services regardless of whether they decided to receive treatment so that they could compare retention rates between those who did and did not receive services. They found that students who chose to be in counseling had significantly higher retention rates (79%) as compared with those who did not (65%). Snell et al. (2001) and Vonk and Thyer (1999), on the other hand, used clinical symptom measures to assess effectiveness; however, Snell et al.'s assessment was between pretest and follow-up, whereas Vonk and Thyer's was between pre- and posttreatment. Specifically, Snell et al. sent out follow-up surveys that consisted of items from a computerized intake assessment to clients who had terminated 10 months earlier. They found that 68% of the 158 clients who received a minimum of one session reported a lower level of symptoms at follow-up when compared with their level at intake. The magnitude of their treatment effect

between pretest and 10-month follow-up was  $d = 0.41$  on the basis of reported data. Similarly, Vonk and Thyer, using the global severity index (GSI) of the Symptom Checklist-90-R (SCL-90-R; Derogatis, 1992; Derogatis & Spencer, 1982), calculated their observed pre- and posttreatment effect size at  $d = 0.84$  ( $N = 41$ ) for clients who completed their posttreatment assessment at a planned termination session.

As long as data are provided, the magnitude of the treatment effect at UCCs could certainly be estimated from studies that focused on other research questions. For example, based on reported pre- and posttreatment data, Tracey et al.'s (1999) observed treatment effect size, also calculated with the GSI, was  $d = 0.76$  ( $N = 20$ ). Similarly, observed pre–post effect size from Hogg and Deffenbacher's (1988) comparison between cognitive and interpersonal process therapy was  $d = 2.11$  ( $N = 27$ ) with the Beck Depression Inventory (BDI; Beck & Steer, 1984). Likewise, in one of the clinical feedback studies conducted by Lambert et al. (2001), they reported that their observed pre- and posttreatment effect size was  $d = 1.04$  ( $N = 609$ ).

But what do these numbers mean? How could these effect size estimates serve as indicators of treatment effectiveness? In order to interpret the absolute magnitude of these numbers, we need an external indicator that would allow us to assess their magnitude rather than relative indicators, notably comparison groups included within each study. In other words, although incorporating comparison groups would allow us to interpret the relative magnitude of the observed treatment effects between groups, these comparative designs provide no insight into how the absolute magnitude should be interpreted. Therefore, unless the effect size estimates obtained from UCCs are compared against some external criterion, lack of interpretability of these numbers renders them ineffectual as indicators of effectiveness.

One method that could provide interpretability of the treatment effect size estimates is benchmarking (e.g., Merrill, Tolbert, & Wade, 2003; Wade, Treat, & Stuart, 1998; Weersing & Weisz, 2002). In the context of evaluating the effectiveness of psychotherapy treatment, Wade et al. (1998) succinctly summarized benchmarking as follows: “[I]n essence, we use the magnitude of change obtained in efficacy studies as a benchmark against which to judge the magnitude of change in service clinic settings” (p. 231). Therefore, they used two clinical trials—notably Barlow, Craske, Cerny, and Klosko (1989) and Telch et al. (1993)—as their benchmarks to assess the outcomes of individual and group cognitive–behavioral therapy (CBT) treatments provided at a community mental health center (CMHC). Wade et al. concluded that their CMHC treatment outcomes and the benchmarks were similar, although they did not statistically compare their data with the benchmarks. Using a similar design and identical methodology in the same CMHC setting, Merrill et al. (2003) concluded that the effectiveness of CBT provided to clients with major depression compared favorably with the clinical trials benchmark.

Weersing and Weisz (2002) significantly improved on Wade et al.'s (1998) methodology by incorporating meta-analysis in constructing their benchmarks. In their investigation of effectiveness of treatment provided to youth with depressive symptoms in CMHCs in the greater Los Angeles area, they established a “research standard of care” (p. 301) by meta-analytically aggregating all psychotherapy clinical trials for youth with depression. They found that, compared against this benchmark, the clinical trajec-

tory of youth treated in these CMHCs clearly resembled that of wait-list control clients in clinical trials.

Further building on Weersing and Weisz's (2002) methodology, Serlin, Wampold, and colleagues proposed a statistical analysis to investigate effectiveness and conducted a benchmarking study on psychotherapy treatment outcome provided in a managed care environment (HMO; Minami, Serlin, Wampold, Kircher, & Brown, 2008; Minami, Wampold, et al., 2008). As did Weersing and Weisz, they constructed benchmarks by meta-analytically aggregating clinical trials of psychotherapy treatment for adult depression (Minami, Wampold, Serlin, Kircher, & Brown, 2007). Their statistical analyses suggested that clients who received psychotherapy treatment in HMOs are likely receiving effective treatment as compared with clinical trials, regardless of whether the clients were on antidepressant medication.

As promising as benchmarking might be to assess the absolute treatment effect of UCCs, benchmarking is not without its problems. Comparison against clinical trials, regardless of their internal validity and the reliability of their findings, is far from ideal because UCCs and clinical trials differ drastically (Minami & Wampold, 2008; Minami, Wampold, et al., 2008; Nathan, Stuart, & Dolan, 2000; Rounsaville, O'Malley, Foley, & Weissman, 1988; Seligman, 1995; Wampold, 1997, 2001; Westen & Morrison, 2001; Westen, Novotny, & Thompson-Brenner, 2004). Specifically, in contrast to most UCCs, clinical trials randomly assign clients, have a set number of sessions (typically between 12 and 20), train therapists with a specific treatment manual, provide supervision by experts in the treatment under investigation, and exclude clients with comorbid conditions. Furthermore, the work environment of the therapists differs significantly. However, because (a) reliable benchmarks of psychotherapy effectiveness in the real world, including UCCs, currently do not exist and (b) psychotherapy efficacy as evidenced in clinical trials have for some time been the "gold standard" (Seligman, 1995, p. 966; see also Chambless & Hollon, 1998; Goldfried & Wolfe, 1998; Kraemer, Wilson, Fairburn, & Agras, 2002), it is still reasonable to compare observed effect sizes obtained in UCCs against clinical trials despite these limitations. Benchmarks constructed from efficacy observed in clinical trials, albeit nowhere near perfect, are still the best we currently have.

In the current study, we attempted to evaluate the effectiveness of counseling services provided at a UCC by benchmarking its observed pre-post effect size estimate against treatment efficacy benchmarks constructed from clinical trials as reported in Minami et al. (2007). Specifically, we hypothesized that the magnitude of the observed treatment effect would be clinically equivalent to treatment efficacy observed in clinical trials.

## Method

### Setting

For this study, archival clinical data from a large western public university's UCC (hereafter referred to as the center) was used. In addition to providing direct clinical services, the center serves the university community by providing outreach and consultation. The center also heavily emphasizes its role in training both master's and doctoral students in psychology and social work to be competent in providing individual and group counseling, psychological

assessment, outreach, and consultation. Additionally, interns from both social work and psychology are accepted every year, and the center's psychology predoctoral internship program is accredited by the American Psychological Association.

With regard to clinical services, the center provides more individual than group counseling. Although the majority of the clients are students, clinical services are also offered to faculty and staff. Students are uniformly charged \$10 per individual session and \$5 per group session except for at intake, which is free of charge. Faculty and staff are charged on a sliding scale according to their reported income. Individual therapy is generally limited to 12 sessions per year, although this is a rather flexible limit, allowing the therapist and client to mutually decide on treatment duration and frequency. The center enrolls clients into individual therapy far more than group therapy, even though there are no session limits posed on group therapy. Pharmacotherapy is also offered at the center through a part-time psychiatrist and psychiatric residents.

### Clients

A total of 6,099 adult clients attended a total of 38,360 sessions at the center between August 5, 1999, and May 31, 2007. Because client demographics data have not been reliably entered until fairly recently, more than half of the data were missing or invalid; however, of the 2,691 clients (44.1%) whose demographic data were deemed reliable, 60% reported being female and 40% reported being male; 55% reported being single, 37% reported being partnered and/or married, 7% reported being separated or divorced, and 1% reported other; mean age was  $M = 27.39$  years ( $SD = 7.92$ ;  $Mdn = 25$ ; range = 18–70)<sup>1</sup>; reported racial/ethnic identifications were 5% Hispanic (non-White), 4% Asian American, 1% Black, 1% Native American, and 89% White. Diagnoses or other information regarding the nature of the clients' presenting concerns also have not been routinely collected at the center and thus were not available for any of the sessions in the database. However, the center's most recent annual survey (conducted for preparing administrative reports) indicated that over a 6-month period, depression has consistently been the highest presenting concern reported by clients (59%), followed by anxiety (56%), stress (48%), and academic issues (41%). In addition, during clinical staff meetings, clients that are (a) actively suicidal with comorbid complicating factors, (b) likely struggling with a personality disorder, and/or (c) deemed to stay longer than the 12-session limit are actively referred out to mental health services available in the community due to limitations in resources. Clients are considered to potentially be long-term (i.e., exceed the 12-session limit) at the center if extensive psychiatric history is

<sup>1</sup> For unidentifiable reasons, these values are likely biased upward, as our most recent and most reliable data of 530 clients seen between July 1, 2007, and October 13, 2008, indicate  $M = 25.51$  years ( $SD = 7.04$ ,  $Mdn = 24$ , range = 18–63). Regardless, the relatively high average age of the clients at the center is likely due to two factors: (a) some of the clients are staff and faculty (albeit data are not available on their status) and (b) up to 60% of the students at the university are estimated to be members of the Church of Jesus Christ of Latter-Day Saints, whose teachings require men to complete a 2-year mission during early adulthood (women are also encouraged, but not required, to go on a 1 1/2-year mission).

revealed during the intake interview, including numerous psychiatric hospitalizations in the past and/or severe alcohol and/or drug addiction. For the current analysis, the number of clients was further reduced (as described later in the *Data Reduction* section).

### Therapists

During the above 8 years, 191 therapists who conducted an intake or individual counseling were identified in the original database. At the center, therapists included full-time clinical staff consisting of psychologists, clinical social workers, and licensed professional counselors, as well as trainees at various levels (e.g., psychology practicum students, psychology predoctoral interns, social work interns, postdoctoral clinical staff). Under the data use agreement for this study, most therapists' demographic information and other professional data (e.g., race/ethnicity, years in practice) were intentionally kept inaccessible because some of the authors were affiliated with the center. However, during this period, most of the therapists were White, the majority were women, and, at any year, trainees provided a roughly equal amount of direct clinical hours as did clinical staff in total volume. In a survey conducted for a different study, most of the staff indicated that their theoretical orientation was integrative or that they practiced from multiple theoretical orientations and perspectives, including psychodynamic, cognitive-behavioral, interpersonal, humanistic, existential, feminist, and multicultural.

### Measure of Treatment Outcome

Treatment outcome was assessed with the Outcome Questionnaire-45.2 (OQ-45; Lambert et al., 2004). The OQ-45 is a 5-point Likert-type scale ranging from 0 (*never*) to 4 (*almost always*). Each of the 45 items are assigned to one of three subscales, namely Symptom Distress, Interpersonal Relations, and Social Role Performance. Nine of the 45 items are reverse-scored. The total and each subscale's scores are derived by adding each of the items without weighting. The actual items are provided in Vermeersch et al. (2004).

The OQ-45 has been used in 40 UCCs nationwide (Vermeersch et al., 2004). In the authors' evaluation, the total and subscale scores as well as 34 of the 45 items had significant sensitivity to change. With UCC samples, Umphress, Lambert, Smart, Barlow, and Clouse (1997) have established the concurrent validity of the OQ-45 using the Inventory of Interpersonal Problems ( $r = .66$ ; Horowitz, Rosenberg, Baer, Ureno, & Villasenor, 1988), the Social Adjustment Scale ( $r = .79$ ; Weissman & Bothwell, 1976), and the SCL-90-R ( $r = .78$ ; Derogatis, 1992). For reliability estimates, Lambert et al. (2004) reported  $r = .82$  for a 3-week test-retest and Cronbach's alpha of .93.

Lambert et al. (2004, 2001) determined a clinical cutoff to distinguish between the nonclinical community ( $N = 1,353$ ) and clinical population ( $N = 1,476$ ) using the following formula given by Jacobson and Truax's (1991):

$$c = \frac{SD_1M_2 + SD_2M_1}{SD_1 + SD_2}, \quad (1)$$

where  $M_1$  and  $SD_1$  are the mean and standard deviation of the community sample, and  $M_2$  and  $SD_2$  are those of the clinical sample. Accordingly, they reported that symptoms of clients who

scored at the cutoff score of 63 or below were more similar to those of a nonclinical community sample.

### Collection of Archival Data

This study analyzed archival data from the center's records collected between August 5, 1999, and May 31, 2007. Since 1999, all center clients have been requested to complete the OQ-45 (i.e., self-report) prior to seeing the therapist at every visit. Specifically, at intake, clients are instructed to arrive approximately 30 min prior to their session to fill out paperwork, consent forms, and their first OQ-45. Clients are encouraged to arrive early enough for their subsequent sessions so that they have time to complete the OQ-45 prior to seeing their therapist. Although no formal data are available, most clients complete the OQ-45 within 5 min, and seldom do clients refuse to complete it. Upon completion of the OQ-45, the clients hand it to their therapist, who then typically inspects particular items such as those indicating suicidal ideation and substance use. The clients' responses to the OQ-45 are then entered into the center's database. Therefore, unless the therapists calculate the total score themselves (which is extremely rare), they do not know their clients' total score on the OQ-45 until it is entered into the database.

### Data Reduction

For the current study, data on individual sessions were organized into treatment cases. Whereas an individual case could easily be defined in clinical trials because research protocols determine the first and last treatment sessions, this does not apply in natural clinical settings. Therefore, it was determined that when clients had not returned to the center for over 90 days, the last assessment that the client filled out was considered their posttreatment assessment.<sup>2</sup> If the client returned to the center after a 90-day gap, the first assessment after the gap was considered to be the intake assessment for a subsequent treatment case. In addition, the following inclusion criteria were applied: (a) only cases with two or more recorded sessions were included in order to calculate an effect size, (b) only clients with initial OQ-45 scores above the clinical cutoff score (i.e., 63) were included to best match level of severity with the clinical trials benchmarks, and (c) only one case per client was included to maintain independence of observations at the client level. When there was more than one case per client in the data, the first treatment case was selected for inclusion.

Specifically, the 38,360 sessions from 6,099 clients were first organized into 7,650 (100%) cases. Among these, 3,002 (39.24%) cases were intake only (e.g., clients chose not to return to counseling after their intake). The remaining 4,648 cases belonged to 3,800 clients, and therefore 848 cases were taken out to maintain data independence at the client level. Of the 3,800 cases, 1,128 (24.68%) cases did not meet the initial severity criterion and were

<sup>2</sup> Although the number of days used to determine the break between cases is arguably arbitrary, this value was used because a 3-month summer vacation was a commonly observed break among students. University policy also dictates that students who are not enrolled full-time during the summer cannot be admitted to the center for individual therapy, whereas this policy is not in effect during the winter break and many clients do tend to continue therapy during this period.

thus excluded. Therefore, the above data reduction procedure resulted in 2,672 (34.93%) cases of counseling (i.e., 2,672 clients) seen by 148 different therapists.

### Benchmarking

The benchmarking procedure used in this study was adapted from Wampold, Serlin, and colleagues (Minami, Serlin, et al., 2008; Minami, Wampold, et al., 2008; Minami et al., 2007). In general, their strategy involved three steps: (a) constructing the benchmark(s), (b) calculating a pre–post treatment effect size estimate observed in the clinical setting, and (c) statistically comparing the observed effect size estimate against the constructed benchmarks.

**Benchmarks selection.** Rather than developing new benchmarks, we adapted those from Minami et al. (2007), in which benchmarks were constructed by meta-analytically aggregating standardized pre–post mean change scores of psychotherapy treatment and wait-list control conditions in published clinical trials of adult major depression. In an ideal treatment effectiveness investigation, the outcome measure between the benchmark and the UCC data is identical. Although clinical trials heavily favor certain clinical outcome measures such as the BDI, SCL-90-R, and Hamilton Rating Scale for Depression (HRSD; Hamilton, 1960, 1967), many clinical service settings (e.g., UCCs) are less likely to routinely utilize these measures due to resource limitations (e.g., time, cost) and/or practicality (e.g., length- or diagnosis-based). In these situations, although imperfect, outcome measures used in the benchmarks and for assessing treatment effectiveness should be matched on the basis of two criteria (Lambert, Hatch, Kingston, & Edwards, 1986; Minami et al., 2007; Smith, Glass, & Miller, 1980). The first criterion, reactivity, is determined on the basis of who reported the symptoms, notably (a) a clinician other than the treating therapist (i.e., high reactivity) or (b) the client (i.e., low reactivity). The second criterion, specificity, refers to whether the measure was designed to assess (a) clinical symptoms of a specific diagnosis (i.e., high specificity; e.g., BDI, HRSD) or (b) general clinical symptoms (low specificity; e.g., SCL-90-R). Meta analyses have consistently revealed that measures that are higher in reactivity and specificity result in larger effect sizes (Lambert et al., 1986; Minami et al., 2007; Smith et al., 1980). For example, in Minami et al., the largest effect size was observed with the HRSD (i.e., high reactivity and specificity;  $d = 2.43$ ), followed by the BDI (i.e., low reactivity but high specificity;  $d = 1.71$ ) and lastly the composite measure comprised of scales that are low on both reactivity and specificity ( $d = 0.80$ ). Thus, in cases where the benchmark and the outcome measure used to assess effectiveness cannot be identical, the measures should be matched on both reactivity and specificity.

Given the typical but less than ideal situation in the current investigation, benchmarks aggregating measures low on both reactivity and specificity (LR–LS) were selected so as to match those of the OQ-45, which is also low in reactivity and specificity. In particular, three benchmarks were adapted to assess the effectiveness of treatment at the center: intent-to-treat (ITT;  $d_{E(ITT)} = 0.795$ ), completers ( $d_{E(C)} = 0.932$ ), and wait-list control ( $d_{WLC} = 0.149$ ). In clinical trials, *ITT* refers to all participants (and consequently their data) who are accepted into a study, whereas *completers* refers to those within ITT who continue in treatment until

the agreed-upon termination. Thus, because what distinguishes the ITT and completer groups is whether they include participants who withdraw prematurely, observed treatment effect sizes tend to be lower for the ITT participants as compared with those for the completers (Minami et al., 2007). Therefore, comparison against the ITT benchmark would indicate how large the treatment outcome is in comparison to all clients who were accepted into the clinical trials, whereas the comparison against the completer benchmark would assess the magnitude of the treatment outcome in comparison to all clients who completed the treatments in the clinical trials. The wait-list control benchmark is constructed using clinical trials data from clients who were randomized into wait-list control conditions (Posternak & Miller, 2001). Therefore, comparison against the wait-list control benchmark assesses whether services improved clients' psychological distress beyond what was observed in natural symptom remission.

**Effect size calculation.** Basic meta-analytic procedures were used to calculate the observed effect size in units of standardized pre–post mean change (Becker, 1988; Hedges & Olkin, 1985; Morris, 2000). The standard deviation of the intake score was used for the standardization because, unlike the pooled standard deviation, it is not influenced by repeated testing and treatment effect (Becker, 1988; Morris, 2000). Approximation given by Morris (2000, p. 19, formula 9) rather than other popular approximations was used to calculate the variance (i.e., squared standard error) of the unbiased estimator because of the higher degree of accuracy.

**Statistical analysis.** Statistically comparing the observed treatment effect size against the benchmarks creates a dilemma in studies with high statistical power (Serlin & Lapsley, 1985, 1993). In other words, with enough participants, any difference against 0, however small, can result in statistical significance. Therefore, it becomes necessary to determine a priori the magnitude of difference between the benchmark and the effect size estimate that could be considered clinically trivial. Once this margin is determined, a range-null (as opposed to the traditional point-null) hypothesis test is conducted (Serlin & Lapsley, 1985, 1993). Specifically, in a one-tailed investigation (with  $\alpha = .05$ ) interested in assessing whether the observed treatment effect is at minimum within the benchmark minus the margin, the range-null and alternative hypotheses with an a priori margin  $d_{\Delta}$  that is considered clinically trivial are, respectively,

$$H_0: \delta_B - \delta_T \geq d_{\Delta} \text{ and} \quad (2)$$

$$H_1: \delta_B - \delta_T < d_{\Delta}, \quad (3)$$

where  $\delta_T$  is the population treatment effect size and  $\delta_B$  is the benchmark to surpass. The test statistic follows a noncentral  $t$  distribution with  $\nu = N - 1$  degrees of freedom and a noncentrality parameter  $\lambda = \sqrt{N}(\delta_B - d_{\Delta})$ . Thus, the critical value  $d_{CV}$  that the observed effect size needs to surpass, in order to claim that the treatment is as effective as the benchmark, is

$$d_{CV} = t_{\nu, \lambda; .95} / \sqrt{N}, \quad (4)$$

where  $t_{\nu, \lambda; .95}$  is the 95th percentile value of the above noncentral  $t$  distribution.

Following Minami, Wampold, et al. (2008), we considered 10% of the benchmarks to be clinically comparable. In other words, if statistical analyses showed that the effect size estimate was within

the efficacy benchmarks minus 10% (i.e.,  $d_{E(ITT)90\%} = 0.715$  and  $d_{E(C)90\%} = 0.839$ ), then it was deemed that the magnitude of the effect size estimate was close enough to the respective benchmarks. Benchmarking against the wait-list control benchmark was statistically identical to the above except for using the 10% margin in the opposite direction. In other words, if the effect size estimate did not exceed the wait-list control benchmark plus 10% (i.e.,  $d_{(WLC)110\%} = 0.163$ ), it was deemed clinically comparable to the wait-list control benchmark. Therefore, if the treatment effect size estimate did not significantly exceed this value, treatment was considered practically equivalent to a wait-list control condition in clinical trials.

*Relative magnitude of effect size estimate against benchmarks.* We were interested in estimating, in addition to the traditional reject versus fail to reject statistical analysis, the relative magnitude of the effect size estimate derived from the center as compared against the efficacy benchmarks. In other words, the question that we attempted to answer was, How did the treatment of interest fare against treatments offered in clinical trials? Therefore, an index of relative magnitude (RM) was calculated so as to illustrate, with a 95% confidence, how large the effect size estimate was in comparison to the efficacy benchmarks. Specifically, given  $\delta_E$ , the true population efficacy benchmark estimated from clinical trials, and  $d_{UCC}$ , the effect size estimate calculated using data from UCCs with  $N$  participants in the data,

$$RM = \frac{\lambda_{UCC}}{\sqrt{N}\delta_E}. \quad (5)$$

When  $\nu (= N - 1)$  and Type I error rate is  $\alpha = .05$ ,  $\lambda_{UCC}$  is the noncentrality parameter when  $t_{UCC} (= \sqrt{N}d_{UCC})$  equals the noncentral  $t$  critical value  $t_{\nu, \lambda(UCC), \alpha}$ . The RM index could easily be interpreted as percentages when multiplied by 100; for example, if  $RM = 0.9$ , the magnitude of the treatment effect was at least 90% as compared with treatments in clinical trials.

### *Assessing the Effect of Nonindependence of Observations at the Therapist Level*

An alternative calculation of the observed effect size estimate was conducted to assess the effect of data nonindependence at the therapist level. Specifically, rather than weighting the cases equally at the level of the client, we calculated the alternative effect size estimate by weighting the cases inversely to the number of clients that the therapist saw. For example, if Therapist A saw only one case (Case A) but Therapist B saw two cases (Cases B1 and B2), then Case A was weighted at 1 and Cases B1 and B2 were each weighted at 1/2 when aggregating.

A further complication was posed due to the clinical team approach taken at the center, where the majority of clients are reassigned to a different therapist than the one at intake. In addition, because trainees reassign their clients to different therapists when they complete their training, clients who receive more sessions have higher likelihoods of seeing more than one therapist over the course of their treatment. Consequently, of the 2,672 clients, 892 (33.4%) clients saw only one therapist, 1,443 (54.0%) clients saw two therapists, 245 (9.2%) clients saw three therapists, and the remaining 92 (3.4%) clients saw four or more therapists. Because the interest was in obtaining independence at the therapist

level, the alternative effect size estimate analysis was conducted only with the 892 cases where clients saw only one therapist. The 892 cases were then further divided into three categories depending on therapists' training level, namely staff, interns, and other trainees. Staff members consist of therapists who have completed their graduate training in either psychology or social work, and all are either licensed or in the process of obtaining licensure. Interns include predoctoral psychology interns, master's of social work interns, and occasionally, master's of counseling interns in a Licensed Professional Counselor program. Other trainees are generally doctoral practicum students who are in their first clinical placement, although some students enter their doctoral program with master's degrees that included clinical work requirements.

### *Explorations of Clinical Trends*

*Effect of client and therapist demographics.* Potential effect of client and therapist demographic variables on observed effect sizes and number of sessions was explored. Prior to analyses, the effects of initial severity and total number of sessions were removed from the raw observed effect sizes (i.e., total residualized effect size;  $R_{ES(T)}$ ) because we were interested in knowing whether demographic variables affected treatment outcomes under clinically comparable conditions. Explored client demographics were age, race/ethnicity, gender, and relationship status. Racial/ethnic categories provided in the demographics sheet were Asian American, Black/African American, Hispanic/Latino, Native American, and White/Caucasian American. Possible effect of therapists' gender as well as match in gender between therapists and clients were also explored. The total number of sessions,  $N_S$ , unlike the pre-post effect size, was not residualized, because the focus of the analysis was whether demographic variables were associated with differential resource allocation. The same demographic variables were explored with  $N_S$ .

*Session frequency and treatment outcome.* Another question of interest was whether having higher frequency of sessions leads to larger treatment outcome. Therefore, correlation between treatment frequency (i.e., total number of sessions divided by the number of days in treatment) and  $R_{ES(T)}$  was explored.

*OQ-45 structure and item sensitivity to change.* To explore the items on the OQ-45 as they related to treatment outcome, we took these three general steps: (a) calculated the residualized effect sizes for each item  $i$  (i.e.,  $R_{ES(i)}$ ), (b) conducted factor analysis on the  $R_{ES(i)}$ , and (c) investigated the magnitude of correlations between raw intake item scores,  $R_{ES(T)}$ , and  $N_S$  in light of the derived factors. First, as with  $R_{ES(T)}$ , we residualized item-level effect sizes, taking into consideration initial severity and number of sessions (i.e.,  $R_{ES[1]}$  through  $R_{ES[45]}$ ) rather than using the raw pre-post change on each of the items. Then, an alpha factor analysis with oblique varimax rotation was conducted on the  $R_{ES(i)}$ s. Alpha factor analysis was chosen so as to maximize generalizability with regard to the factors rather than the participants (Kaiser & Caffrey, 1965); oblique varimax rotation was chosen so as to take into consideration the nonzero interfactor correlations while maximizing the variance of the squared factor loadings (Kaiser, 1958). Factor analysis was conducted on  $R_{ES(i)}$  rather than on raw item scores obtained at a single time point because we were interested in identifying items that change together rather than those that were elevated together at one point in

time. This factor analysis was motivated by the center’s therapists’ heuristic experiences that (a) certain items on the OQ-45, such as those indicating problematic substance use, were likely indicators of clients who would require longer treatment and (b) certain items did not intuitively fit into the three original subscales (e.g., the three items indicating substance use are spread out in each of the three subscales). Therefore, we freely explored the structure by setting the cutoff eigenvalue to 1 (Kaiser, 1960).

After conducting the factor analysis, we analyzed correlations between intake item scores and  $R_{ES(T)}$  as well as between intake item scores and  $N_S$  using the Kruskal–Wallis test (Kruskal & Wallis, 1952) with the obtained factors as the categories. We also conducted this analysis using the original three subscales.

### Results

#### Benchmarking

In the center’s data, the average intake and last scores of the  $N = 2,672$  clients were  $M_{pre} = 87.44$  ( $SD_{pre} = 16.44$ ) and  $M_{post} = 71.39$  ( $SD_{post} = 22.16$ ), respectively, resulting in an observed standardized pre–post mean change of  $d = 0.9755$  ( $SE = 0.0240$ ; see Table 1). Assuming the applicability of a normal distribution, the magnitude of the observed effect size estimate indicated that approximately 83.53% of the clients who receive treatment at the center are likely to be clinically less symptomatic as compared with the average client at intake who does not receive treatment; in reality (i.e., this database), the actual percentage of clients who had less severe clinical symptoms at their last session compared with the average severity at intake was 78.18%. The average number of sessions was 6.84 ( $SD = 8.72$ ;  $Mdn = 4$ ) over a span of an average of 88.52 ( $SD = 106.12$ ;  $Mdn = 56$ ) days in treatment. Therefore, on average, clients received sessions on a biweekly basis (i.e.,  $88.52/6.84 = 12.94$  on the basis of means;  $56/4 = 14$  on the basis of medians). As with other clinical data, the distribution of these variables was positively skewed; 75% of the clients had nine sessions or fewer and ended treatment within 115 days.

The magnitude of the center’s effect size estimate was significantly larger than both the ITT and completer treatment efficacy benchmarks as well as the wait-list control benchmark (see Table 2). Specifically, the effect size estimate exceeded the critical value  $d_{CV} = 0.8766$  on the basis of the completer treatment efficacy benchmark, demonstrating that the magnitude of treatment effect was at least as large as the treatment efficacy of completer samples observed in clinical trials. The magnitude of effect as compared with the completer benchmark (i.e.,  $RM = 1.0047$ ) indicated that, with a 95% confidence, the true treatment effect size at the center was likely at minimum 100% as compared with completers in clinical trials. Similarly, the estimated effect exceeded  $d_{CV} = 0.7513$ , which was the magnitude of effect necessary to claim at

least equivalence with ITT treatment efficacy observed in clinical trials. As compared with the ITT benchmark,  $RM = 1.1789$  indicated that the true magnitude of effect was at minimum 118%, again with a 95% confidence. Consequently, with the estimated effect size exceeding both treatment benchmarks, effectiveness was well beyond natural remission, at over six times its magnitude.

#### Alternatively Weighted Effect Size Estimates

Because the data were not independent at the level of therapists, alternative effect size estimates were calculated by weighting each case relative to the caseload of the therapist. To attain complete independence at the therapist level, cases were first categorized on the basis of the number of therapists involved (see Table 3), and then cases with only one therapist were selected for the analysis (i.e.,  $N = 892$  [33.4%],  $d = 0.796$ ). The data for these 892 clients were then aggregated within therapists and weighted inversely to the therapists’ caseload. Contrary to expectations, interns and other trainees had observed effect sizes that were significantly larger than those of staff (both  $ps < .001$ ). The difference could not be fully explained by the fewer number of sessions (as indicated by  $d/N_S$ ; see Table 4). However, reasons for differences in treatment outcome cannot simply be ascribed to therapists’ training level because clients are actively reassigned to therapists on the basis of the clients’ clinical profiles.

#### OQ-45 Clinical Trends

*Effect of client and therapist demographics.* For categorical demographic variables, the Kruskal–Wallis test was used to investigate their effect on the residualized effect size estimate  $R_{ES(T)}$  and total number of sessions  $N_S$ ; for client age, Spearman’s rho ( $\rho$ ) was used. For both  $R_{ES(T)}$  and  $N_S$ , no significant differences were observed on the basis of client race/ethnicity, client gender, therapist gender, or gender match between client and therapist ( $n = 559\sim 605$ ,  $p = .130\sim .948$ ). However, clients’ reported relationship status was significantly related to treatment outcome ( $H = 14.80$ ,  $df = 3$ ,  $p = .002$ ). Specifically, clients who reported that they were either partnered or married had significantly better outcomes than did those who reported their status as separated or divorced (average rank [AR] = 245.5 [of 559] against AR = 323.7, respectively), after taking into consideration their initial severity and number of sessions. The difference in these average ranks was associated with a sizeable average difference of 5.8 points on the OQ-45. Although client age also significantly correlated with number of sessions ( $\rho = .087$ ,  $n = 551$ ,  $p = .041$ ), the percentage of variance explained was only 0.76%.

*Session frequency and treatment outcome.* The correlation between session frequency (i.e., the total number of sessions divided

Table 1  
Effect Size Estimates From University Counseling Center Data

$N_C$	$N_T$	$M_{pre}$ (SD)	$M_{post}$ (SD)	$r_{pre-post}$	$d$	SE	$N_S$ (SD)	$N_D$ (SD)	$d/N_S$
2,672	148	87.44 (16.44)	71.39 (22.16)	.4699	0.9755	0.0240	6.84 (8.72)	88.52 (106.12)	0.143

Note.  $N_C$  = number of cases;  $N_T$  = number of therapists;  $N_S$  = number of sessions;  $N_D$  = number of days in treatment;  $d/N_S$  = effect size estimate divided by the number of sessions.

Table 2  
Benchmarking

<i>d</i>	vs. ITT benchmark			vs. completers benchmark			vs. wait-list control benchmark		
	<i>d<sub>CV</sub></i>	<i>p</i>	RM	<i>d<sub>CV</sub></i>	<i>p</i>	RM	<i>d<sub>CV</sub></i>	<i>p</i>	RM
0.9755	0.7513	<.001	1.1789	0.8766	<.001	1.0047	0.1956	<.001	6.3059

Note. *d<sub>CV</sub>* = critical effect size value to attain statistical significance; RM = index of relative magnitude.

by the number of days in treatment) and  $R_{ES(T)}$  was statistically significant ( $r = -.039$ ,  $n = 2,671$ ,  $p = .046$ ) but in the direction that was counterintuitive. The direction of the correlation indicated that higher frequency led to lower overall treatment outcome after taking into consideration the initial severity and overall number of sessions. However, the magnitude of this effect was less than 1/6 of a percent and thus was unlikely to have any practical relevance.

**OQ-45 structure and item sensitivity to change.** The exploratory alpha factor analysis on  $R_{ES(i)}$ s indicated that there were potentially eight factors: anhedonia (Items 3, 13, 20, 21, 24, 31, 43), psychological distress (Items 5, 8, 10, 15, 23, 25, 33, 35, 36, 40, 42), physical distress (Items 2, 9, 27, 29, 34, 41, 45), loss of productivity (Items 12, 22, 28, 38), lack of intimacy (Items 7, 16, 17, 18, 37), problematic substance use (Items 11, 26, 32), interpersonal conflict (Items 1, 19, 3, 39, 44), and stress (Items 4, 6, 14; actual items listed in Vermeersch et al., 2004). Of interest, consistent with heuristic observations at the center, the three items that indicated problematic substance use formed their own factor.

The Kruskal–Wallis test on the average ranks of correlations between raw intake item score and  $R_{ES(T)}$  was significant ( $H = 16.16$ ,  $df = 7$ ,  $p = .024$ ), indicating that elevations on the eight factors at intake differentially correlated with the residual treatment effect after taking into consideration overall initial severity and total number of sessions. Especially the three items indicating problematic substance use contributed significantly to smaller pre-post effect sizes (AR = 41.0 [out of 45]). After inspecting the residuals of these three items, roughly three distinct groups were identified. The 2,247 cases with the sum of the three items totaling 2 or less (out of a possible 12) were, on average, 0.6 points lower in overall symptom distress at end of treatment than what could be expected. In contrast, the 337 cases with a sum of the three items between 3 and 5 were on average 2.2 points higher in overall symptom distress; further, the 88 cases with a sum of 6 or more resulted in an average of 6.8 points higher. Given the average raw score difference between pre- and posttreatment of approximately 16 points, the treatment effect for these 88 cases was less than 60% of what could be expected on average. Elevation at intake on items

indicating physical distress (AR = 33.4) and interpersonal conflict (AR = 28.0) also led to smaller effect sizes, whereas items indicating loss of productivity (AR = 13.3) and psychological distress (AR = 17.0) resulted in larger pre-post effect sizes. The same analysis with the original three subscales (i.e., Symptom Distress, Interpersonal Relations, and Social Role Performance) was not statistically significant ( $H = 0.96$ ,  $df = 7$ ,  $p = .618$ ). Specific item of interest, notably one item regarding suicidal ideation, had no evidence of leading to poorer outcome ( $r < .001$ ,  $p = .976$ ).

The Kruskal–Wallis test on the average ranks of the correlations between raw intake item scores and  $N_S$  was also significant ( $H = 24.29$ ,  $df = 7$ ,  $p = .001$ ), indicating that elevations on the eight factors differentially correlated with the total number of sessions. The five items indicating lack of intimacy (AR = 34.4) contributed the most to increased number of sessions; linearly estimated difference in number of sessions between the lowest possible total score on these five items (i.e., 0) and the highest (i.e., 20) was approximately 2.8 sessions. Anhedonia (AR = 31.9) was also related to increase in sessions, with a linearly estimated difference of 2.6 sessions between the highest and lowest total scores on these seven items. Substance issues (AR = 7.0), loss of productivity (AR = 8.0), and stress (AR = 10.0) were related to fewer number of sessions. The same analysis with the original three subscales was also significant ( $H = 9.15$ ,  $df = 2$ ,  $p = .010$ ). The 12 items making up Interpersonal Relations (AR = 27.5), which include all five items indicating lack of intimacy, significantly contributed to increased number of sessions. Suicidal ideation did not lead to increased number of sessions ( $r = .031$ ,  $p = .112$ ).

## Discussion

The absolute magnitude of the effectiveness of counseling services at UCCs has rarely been investigated despite their being one of the major mental health services providers. The question of how effective treatments provided at UCCs are is of interest to therapists, researchers, administrators, and above all, clients; thus, this

Table 3  
Effect Size Estimates on the Basis of Number of Therapists

$N_T$	$N_C$	$M_{pre} (SD)$	$M_{post} (SD)$	$r_{pre-post}$	<i>d</i>	<i>SE</i>	$N_S (SD)$	$N_D (SD)$	$d/N_S$
1	892	88.03 (17.19)	74.33 (21.70)	.5396	0.7959	0.0372	4.49 (4.96)	63.58 (77.59)	0.177
2	1,443	86.34 (15.70)	70.01 (21.74)	.4498	1.0396	0.0337	6.28 (7.07)	82.26 (94.39)	0.165
3	245	90.19 (17.00)	69.61 (24.23)	.4140	1.2071	0.0884	12.24 (10.29)	150.84 (117.03)	0.099
4+	92	91.53 (17.46)	69.25 (25.20)	.3696	1.2654	0.1510	24.03 (16.51)	262.53 (170.35)	0.061

Note. Effect size estimates are not weighted on the basis of the number of cases each therapist had.  $N_T$  = number of therapists;  $N_C$  = number of cases;  $N_S$  = number of sessions;  $N_D$  = number of days in treatment;  $d/N_S$  = effect size estimate divided by the number of sessions.



Table 4  
Effect Size Estimates by Training Level on Single-Therapist Cases

Training level	$N_C$	$N_T$	$M_{pre}$ (SD)	$M_{post}$ (SD)	$r_{pre-post}$	$d$	$SE$	$N_S$ (SD)	$N_D$ (SD)	$d/N_S$
All (weighted)	892	122	88.19 (17.19)	73.49 (21.70)	.5396	0.8543	0.0380	4.49 (4.96)	63.58 (77.59)	0.190
Staff	481	30	87.14 (17.51)	77.18 (21.93)	.5399	0.5681 <sub>a,b</sub>	0.0475	4.35 (5.03)	67.80 (91.78)	0.131
Interns	312	61	90.04 (16.52)	73.53 (21.21)	.5103	0.9970 <sub>a</sub>	0.0690	4.80 (5.09)	59.62 (59.62)	0.208
Other trainees	85	41	85.05 (17.55)	68.74 (22.81)	.6531	0.9206 <sub>b</sub>	0.1158	4.22 (4.24)	56.69 (52.47)	0.218

*Note.* Effect size estimates are weighted by the inverse of the total number of cases each therapist saw. Effect sizes with the same subscripts were significant at  $p < .001$ .  $SD$ s,  $r$ s,  $d$ s, and  $SE$ s are from the total number of cases within each training level (i.e., not the number of therapists), as they serve as better estimates than do data aggregated by therapists. Total number of cases for all training levels is larger than the combined total cases seen by staff, interns, and other trainees (i.e.,  $481 + 312 + 85 = 878 < 892$ ) because some cases did not have valid therapist IDs to code training levels. The combined total of staff, interns, and other trainees did not equal the number of therapists in the all (weighted) condition (i.e.,  $30 + 61 + 41 = 132 > 122$ ) because some trainees later became interns and/or staff at the center.  $N_C$  = number of cases;  $N_T$  = number of therapists;  $N_S$  = number of sessions;  $N_D$  = number of days in treatment;  $d/N_S$  = effect size estimate divided by the number of sessions.

study attempted to investigate their effectiveness with a benchmarking method.

Analysis of the observed treatment effect size at this center indicates that counseling services delivered to clients with clinically significant distress were very effective. In particular, for clients who returned to the center for at least one additional session after their intake, the magnitude of the treatment effect was likely equivalent to treatments delivered in clinical trials for adult clients with major depression. Evaluation of the observed treatment effectiveness against the wait-list control benchmark suggests that approximately 80% of the clients treated for two or more sessions at this center were likely better off after receiving treatment than is the average client randomized into a wait-list control condition. Therefore, despite differences in clinical and demographic characteristics between the center and clinical trials included in the benchmark, we find it reasonable to conclude that counseling services provided at this center are very effective.

Contrary to common expectations, treatment outcome at the center did not positively correlate with therapists' training level. Specifically, interns had the highest observed pre-post treatment effect sizes, followed by other trainees (such as practicum students), and then by staff. However, because clients are not randomly assigned to therapists, one must take into consideration the active client reassignment that takes place at the center when interpreting these results. Specifically, of the 2,672 cases, approximately two thirds of the clients were reassigned to a therapist other than their intake therapist. At the center, every therapist belongs to one of four clinical teams consisting of both senior staff and trainees. Thus, all clients who complete their intake session are discussed in one of the four team meetings that the intake therapist participates in. Clients who are deemed appropriate for a brief therapy model are then assigned to a therapist on the basis of (a) therapist's availability in light of overall caseload, (b) therapist's interest in working with the client on the basis of the case report presented by the intake therapist, and (c) team leaders' and other senior staffs' comfort with the assignment. These assignments are rarely disputed, as clients who present with multiple issues, and especially those who likely are experiencing problematic substance use, are actively assigned to senior staff, even if it means that the assignment is made to a senior staff who is on a different team (cross-team referral). Other cases where cross-team referrals occur are the rare occasions when (a) clients express interest in a particular therapist or (b) clients request assignment to a therapist

on the basis of their preference of therapists' demographic characteristics (e.g., gender, sexual orientation, race/ethnicity). Therefore, clients whom senior staff keep on their caseload are those who are likely experiencing clinically difficult problems, including substance use.

Exploratory analyses were conducted on several factors that were of clinical interest. Specifically, client and therapist demographic variables such as age, race/ethnicity, gender, and match in gender between clients and therapists have long been of interest in the profession with mixed results (e.g., Cottone, Drucker, & Javier, 2002; Lambert et al., 2006; MacDonald, 1994; Miranda et al., 2006; Zlotnick, Elkin, & Shea, 1998). However, our study indicated that, at least for this center, few demographic variables had any impact on effectiveness. The one variable that did significantly contribute to treatment outcome, namely clients' relationship status, is not surprising on the basis of both common sense and the social psychology literature (e.g., Baumeister & Leary, 1995; Patrick, Knee, Canevello, & Lonsbary, 2007). What is interesting, however, is that despite the obvious importance of intimate relationships on our psychological well-being, the psychotherapy literature has traditionally focused much more on factors that are solely attributed to the client (e.g., gender, race/ethnicity). The magnitude of impact that clients' relationship status had on effectiveness was substantial; the approximate difference of 6 points between clients who were partnered and those who reported as separated corresponds to a magnitude of over one third of the total pre-post effect size. Therefore, with clients who presented themselves to the center and reported their relationship status as separated or divorced, one route toward overall psychological well-being may be to process the loss of their significant relationship and/or to address their need for building new intimate relationships.

Exploratory analyses on the OQ-45 items also revealed results that have significant implications to a training center with a brief treatment model. Specifically, on the basis of clients' responses to the OQ-45, issues related to loss of productivity and stress are conducive to a short-term model because of the relatively positive prognosis and shorter length of sessions. Therefore, clients with these profiles may be better suited for treatment by interns and other trainees. Other client issues with a relatively positive prognosis are psychological distress and anhedonia, although they tend to require longer sessions than do loss of productivity and stress. Interestingly, although both intimacy and interpersonal conflict

can be conceptualized as relational concerns, prognosis and length of treatment differed. Whereas intimacy concerns tended to result in average prognosis after a significantly greater number of sessions, interpersonal conflict (e.g., frequent arguments, disagreements at work/school) tended to result in poorer progress and a small number of sessions.

One of the more interesting findings on the OQ-45 analyses was that complaints of physical distress symptoms tend to lead to a poorer prognosis. Although overall severity has been considered when reassigning clients, no particular attention has been given to elevations on physical distress symptoms. Given the relatively poorer prognosis of clients with these complaints, active referral to psychiatrists and psychiatry residents may be of benefit because physical symptoms not only might signify possible physical illnesses that require the attention of a physician but also might indicate the need for differential diagnoses and treatment, as in the case of atypical depression (e.g., Angst, Gamma, Benazzi, Ajdacic, & Rössler, 2007).

The clinical issue that seemed most difficult to treat—in terms of overall treatment outcome and likelihood of staying in treatment—appeared to be problematic substance use. Effect size estimates of clients with even slight elevations on the three items (total of 3 to 5 out of a possible 12) were approximately 14% lower than for those who reported minimum elevations (total of 0 to 2); with clients who had substantial elevations on these items (total of 6 or more), their effect size estimates were approximately 42% lower than for those with minimum elevations. Although our analysis was exploratory, our results converge with the literature documenting difficulties in treating clients with problematic substance use (e.g., Dutra et al., 2008). Therefore, as has been heuristically considered at the center, it is unlikely that a brief treatment model will serve the needs of clients with problematic substance use.

There are a number of limitations that make this study preliminary. First, the benchmarks adapted from Minami et al. (2007) were not constructed with the OQ-45, which was the outcome measure for the center's data. Although match in reactivity and sensitivity justified the use of LR-LS benchmarks over other benchmarks (i.e., that of the BDI and HRSD), firm conclusions could be drawn only when the instruments were identical. Another major limitation of the benchmarks was that they were constructed with clinical trials of psychotherapy for treating clients who were diagnosed as having major depression on the basis of the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed., text rev.; *DSM-IV-TR*; American Psychiatric Association, 2000). The clients at the center were never diagnosed on the basis of the *DSM-IV-R*, nor were they selected for inclusion on the basis of depressive symptoms. Therefore, major differences in clinical characteristics may exist between clients in clinical trials and the students at the center.

As mentioned earlier, other differences between clinical trials and those at the center—such as client screening and randomization, session limits, therapist training and supervision, and manualization (Nathan et al., 2000; Rounsaville et al., 1988; Seligman, 1995; Wampold, 1997, 2001; Westen & Morrison, 2001; Westen et al., 2004)—may also have influenced the results. In addition, the current study cannot be generalized to other natural clinical settings (e.g., managed care, community mental health) because of significant differences in client population. Moreover, because this

study was conducted at a UCC with a large number of staff and a significant training component, generalizations cannot be made to other types of UCCs (e.g., Stone, Vespia, & Kanz, 2000; Vespia, 2007). Furthermore, because only clients who met the clinical cutoff score and stayed for at least one additional session were included in the analysis, the results of this study cannot be inferred concerning clients who come with less clinical severity or who decide not to continue after their intake session.

It is also important to note that some of the clients at the center take psychotropic medications. Although reliable data on medication use were unavailable for this study, the center's medication management appointment data from July 1, 2006, to June 30, 2007, indicated that of the 911 clients who were seen during the period, at least 138 (15.1%) clients were on psychotropic medication. This is a conservative estimate because students may be prescribed psychotropic medications by their physicians. Given that Minami, Wampold, et al. (2008) reported an increase in treatment effect of  $d = 0.15$  by use of psychotropic medication in a managed care setting, it is possible that the center's observed effect size calculated with only clients who were not on medication could have been as low as  $d = 0.83$  (which is still above the ITT benchmark but below the completer benchmark). Therefore, it is necessary that replications with data that have reliable information on clients' medication use be conducted.

In light of the numerous limitations, our conclusion is that the current study provided preliminary evidence that at least one UCC has been providing solid clinical care. With regard to treatment outcome, there are serious concerns as to whether treatments in natural clinical settings are effective at all (Bickman, 2002; Stone et al., 2000; Weersing & Weisz, 2002; Weisz, Jensen-Doss, & Hawley, 2006). Therefore, we believe that it is crucial to demonstrate the effectiveness of our clinical services and that this study provides one example of how UCCs could do so. However, utilizing this method would require routine clinical assessments to be implemented. Although satisfaction surveys may provide some audiences with useful information in some contexts (e.g., Seligman, 1995), treatment effectiveness is best measured by direct pre-post clinical assessment with psychometrically sound outcome measures (Brock, Green, Reich, & Evans, 1996; Nielsen et al., 2004; Tracey, 1989). We hope that in the near future more UCCs will implement routine outcome assessment and that this will eventually lead to constructing an effectiveness benchmark that better reflects the clients, therapists, and context of various UCCs.

We believe that this study is also a call for more counseling process research conducted in natural clinical settings. Exploratory as they were, and thus not generalizable to other UCCs, our analyses on client demographics and clinical characteristics provide crucial insight into how the center could modify its service structure to attain better treatment outcomes. This suggests again that UCCs could benefit from implementing routine outcome measures.

In conclusion, we hope that researchers and practitioners find it mutually beneficial to collaborate with one another to assess and improve outcomes in natural clinical settings. Expanding the investigations of counseling and psychotherapy beyond the client-therapist dyad or group and into the environmental and cultural structures surrounding these interactions may also significantly benefit clients. Perhaps we have been focusing too much on the

branches and leaves of trees grown in our favorite greenhouses when there is a whole forest out there.

## References

- American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.
- Angst, J., Gamma, A., Benazzi, F., Ajdacic, V., & Rössler, W. (2007). Melancholia and atypical depression in the Zurich study: Epidemiology, clinical characteristics, course, comorbidity and personality. *Acta Psychiatrica Scandinavica*, *115*(suppl. 433), 72–84.
- Barlow, D. H., Craske, M. G., Cerny, J. A., & Klosko, J. S. (1989). Behavioral treatment of panic disorder. *Behavior Therapy*, *20*, 261–282.
- Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, *117*, 497–529.
- Beck, A. T., & Steer, R. A. (1984). Internal consistencies of the original and revised Beck Depression Inventory. *Journal of Clinical Psychology*, *40*, 1365–1367.
- Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology*, *41*, 257–278.
- Bickman, L. (2002). The death of treatment as usual: An excellent first step on a long road. *Clinical Psychology: Science and Practice*, *9*, 195–199.
- Brock, T. C., Green, M. C., Reich, D. A., & Evans, L. M. (1996). The Consumer Reports study of psychotherapy: Invalid is invalid. *American Psychologist*, *51*, 1083.
- Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology*, *66*, 7–18.
- Cottone, J. G., Drucker, P., & Javier, R. A. (2002). Gender differences in psychotherapy dyads: Changes in psychological symptoms and responsiveness to treatment during 3 months of therapy. *Psychotherapy: Theory, Research, Practice, and Training*, *39*, 297–308.
- Davies, D. R., Burlingame, G. M., Johnson, J. E., Gleave, R. L., & Barlow, S. H. (2008). The effects of a feedback intervention on group process and outcome. *Group Dynamics: Theory, Research, and Practice*, *12*, 141–154.
- Derogatis, L. R. (1992). *SCL-90-R administration, scoring, and procedures manual II*. Towson, MD: Clinical Psychometric Research.
- Derogatis, L. R., & Spencer, M. S. (1982). *The Brief Symptom Inventory (BSI): Administration, scoring, and procedures manual I*. Baltimore: Johns Hopkins University School of Medicine, Clinical Psychometrics Research Unit.
- Draper, M. R., Jennings, J., Baron, A., Erdur, O., & Shankar, L. (2002). Time-limited counseling outcome in a nationwide college counseling center sample. *Journal of College Counseling*, *5*, 26–38.
- Dutra, L., Stathopoulou, G., Basden, S. L., Leyro, T. M., Powers, M. B., & Otto, M. W. (2008). A meta-analytic review of psychosocial interventions for substance use disorders. *American Journal of Psychiatry*, *165*, 179–187.
- Erdur, O., Rude, S. S., & Baron, A. (2003). Symptom improvement and length of treatment in ethnically similar and dissimilar client-therapist pairings. *Journal of Counseling Psychology*, *50*, 52–58.
- Gallagher, R. P. (2009). *National survey of counseling center directors*. Alexandria, VA: International Association of Counseling Services.
- Goldfried, M. R., & Wolfe, B. E. (1998). Toward a more clinically valid approach to therapy research. *Journal of Consulting and Clinical Psychology*, *66*, 143–150.
- Hamilton, M. A. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry*, *23*, 56–62.
- Hamilton, M. A. (1967). Development of a rating scale for primary depressive illness. *British Journal of Social and Clinical Psychology*, *6*, 278–296.
- Hayes, J. A. (1997). What does the Brief Symptom Inventory measure in college and university counseling center clients? *Journal of Counseling Psychology*, *44*, 360–367.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.
- Hogg, J. A., & Deffenbacher, J. L. (1988). A comparison of cognitive and interpersonal-process group therapies in the treatment of depression among college students. *Journal of Counseling Psychology*, *35*, 304–310.
- Horowitz, L. M., Rosenberg, S. E., Baer, B. A., Ureno, G., & Villasenor, V. S. (1988). Inventory of interpersonal problems: Psychometric properties and clinical applications. *Journal of Consulting and Clinical Psychology*, *56*, 885–892.
- Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12–19.
- Kahn, J. H., Achter, J. A., & Shambaugh, E. J. (2001). Client distress disclosure, characteristics at intake, and outcome in brief counseling. *Journal of Counseling Psychology*, *48*, 203–211.
- Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*, 187–199.
- Kaiser, H. F. (1960). The application of electronic computers in factor analysis. *Educational and Psychological Measurement*, *20*, 141–151.
- Kaiser, H. F., & Caffrey, J. (1965). Alpha factor analysis. *Psychometrika*, *30*, 1–14.
- Kivlighan, D. M., Jr., McGovern, T. V., & Corazzini, J. G. (1984). Effects of content and timing of structuring interventions on group therapy process and outcome. *Journal of Counseling Psychology*, *31*, 363–370.
- Kraemer, H. C., Wilson, G. T., Fairburn, C. G., & Agras, W. S. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry*, *59*, 877–883.
- Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*, 583–621.
- Lambert, M. J., Hatch, D. R., Kingston, M. D., & Edwards, B. C. (1986). Zung, Beck, and Hamilton Rating Scales as measures of treatment outcome: A meta-analytic comparison. *Journal of Consulting and Clinical Psychology*, *54*, 54–59.
- Lambert, M. J., Morton, J. J., Hatfield, D., Harmon, C., Hamilton, S., Reid, R. C. et al. (2004). *Administration and scoring manual for the OQ-45.2 (Outcome Questionnaire)*. Stevenson, MD: American Professional Credentialing Services.
- Lambert, M. J., Smart, D. W., Campbell, M. P., Hawkins, E. J., Harmon, C., & Slade, K. L. (2006). Psychotherapy outcome, as measured by the OQ-45, in African American, Asian/Pacific Islander, Latino/a, and Native American clients compared with matched Caucasian clients. *Journal of College Student Psychology*, *20*, 17–29.
- Lambert, M. J., Whipple, J. L., Smart, D. W., Vermeersch, D. A., Nielsen, S. L., & Hawkins, E. J. (2001). The effects of providing therapists with feedback on patient progress during psychotherapy: Are outcomes enhanced? *Psychotherapy Research*, *11*, 49–68.
- MacDonald, A. J. (1994). Brief therapy in adult psychiatry. *Journal of Family Therapy*, *16*, 415–426.
- Mallinckrodt, B. (1989). Social support and the effectiveness of group therapy. *Journal of Counseling Psychology*, *36*, 170–175.
- Merrill, K. A., Tolbert, V. E., & Wade, W. A. (2003). Effectiveness of cognitive therapy for depression in a community mental health center: A benchmarking study. *Journal of Consulting and Clinical Psychology*, *71*, 404–409.
- Minami, T., Serlin, R. C., Wampold, B. E., Kircher, J. C., & Brown, G. S. (2008). Using clinical trials to benchmark effects produced in clinical practice. *Quality and Quantity*, *42*, 513–525.
- Minami, T., & Wampold, B. E. (2008). Adult psychotherapy in the real world. In W. B. Walsh (Ed.), *Biennial review of counseling psychology* (Vol. 1, pp. 27–45). New York: Taylor & Francis.

- Minami, T., Wampold, B. E., Serlin, R. C., Hamilton, E. G., Brown, G. S., & Kircher, J. C. (2008). Benchmarking the effectiveness of psychotherapy treatment for adult depression in a managed care environment: A preliminary study. *Journal of Consulting and Clinical Psychology, 76*, 116–124.
- Minami, T., Wampold, B. E., Serlin, R. C., Kircher, J. C., & Brown, G. S. (2007). Benchmarks for psychotherapy efficacy in adult major depression. *Journal of Consulting and Clinical Psychology, 75*, 232–243.
- Miranda, J., Green, B. L., Krupnick, J. L., Chung, J., Siddique, J., Belin, T., & Revicki, D. (2006). One-year outcomes of a randomized clinical trial treating depression in low-income minority women. *Journal of Consulting and Clinical Psychology, 74*, 99–111.
- Morris, S. B. (2000). Distribution of the standardized mean change effect size for meta-analysis on repeated measures. *British Journal of Mathematical and Statistical Psychology, 53*, 17–29.
- Nathan, P. E., Stuart, S. P., & Dolan, S. L. (2000). Research on psychotherapy efficacy and effectiveness: Between Scylla and Charybdis? *Psychological Bulletin, 126*, 964–981.
- Nielsen, S. L., Smart, D. W., Isakson, R. L., Worthen, V. E., Gregersen, A. T., & Lambert, M. J. (2004). The *Consumer Reports* effectiveness score: What did consumers report? *Journal of Counseling Psychology, 51*, 25–37.
- Okiishi, J., Lambert, M. J., Eggett, D., Nielsen, S. L., Dayton, D. D., & Vermeersch, D. A. (2006). An analysis of therapist treatment effects: Toward providing feedback to individual therapists on their clients' psychotherapy outcome. *Journal of Clinical Psychology, 62*, 1157–1172.
- Okiishi, J., Lambert, M. J., Nielsen, S. L., & Ogles, B. M. (2003). Waiting for supershrink: An empirical analysis of therapist effects. *Clinical Psychology and Psychotherapy, 10*, 361–373.
- Patrick, H., Knee, C. R., Canevello, A., & Lonsbary, C. (2007). The role of need fulfillment in relationship functioning and well-being: A self-determination theory perspective. *Journal of Personality and Social Psychology, 92*, 434–457.
- Posternak, M. A., & Miller, I. (2001). Untreated short-term course of major depression: A meta-analysis of outcomes from studies using wait-list control groups. *Journal of Affective Disorders, 66*, 139–146.
- Rounsaville, B. J., O'Malley, S., Foley, S., & Weissman, M. M. (1988). Role of manual-guided training in the conduct and efficacy of interpersonal psychotherapy for depression. *Journal of Consulting and Clinical Psychology, 56*, 681–688.
- Seligman, M. E. P. (1995). The effectiveness of psychotherapy: The *Consumer Reports* Study. *American Psychologist, 50*, 965–974.
- Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist, 40*, 73–83.
- Serlin, R. C., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 199–228). Hillsdale, NJ: Erlbaum.
- Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy*. Baltimore: Johns Hopkins University Press.
- Snell, M. N., Mallinckrodt, B., Hill, R. D., & Lambert, M. J. (2001). Predicting counseling center clients' response to counseling: A 1-year follow-up. *Journal of Counseling Psychology, 48*, 463–473.
- Stone, G. L., Vespia, K. M., & Kanz, J. E. (2000). How good is mental health care on college campuses? *Journal of Counseling Psychology, 47*, 498–510.
- Telch, M. J., Lucas, J. A., Schmidt, N. B., Hanna, H. H., Jaimez, T., & Lucas, R. A. (1993). Group cognitive-behavioral treatment of panic disorder. *Behaviour Research and Therapy, 31*, 279–287.
- Tracey, T. J. (1989). Client and therapist session satisfaction over the course of psychotherapy. *Psychotherapy: Theory, Research, Practice, and Training, 26*, 177–182.
- Tracey, T. J. C., Sherry, P., & Albright, J. M. (1999). The interpersonal process of cognitive-behavioral therapy: An examination of complementarity over the course of treatment. *Journal of Counseling Psychology, 46*, 80–91.
- Umphress, V. J., Lambert, M. J., Smart, D. W., Barlow, S. H., & Clouse, G. (1997). Concurrent and construct validity of the Outcome Questionnaire. *Journal of Psychoeducational Assessment, 15*, 40–55.
- Vermeersch, D. A., Whipple, J. L., Lambert, M. J., Hawkins, E. J., Burchfield, C. M., & Okiishi, J. C. (2004). Outcome Questionnaire: Is it sensitive to changes in counseling center clients? *Journal of Counseling Psychology, 51*, 38–49.
- Vespia, K. M. (2007). A national survey of small college counseling centers: Successes, issues, and challenges. *Journal of College Student Psychotherapy, 22*, 17–40.
- Vonk, M. E., & Thyer, B. A. (1999). Evaluating the effectiveness of short-term treatment at a university counseling center. *Journal of Clinical Psychology, 55*, 1095–1106.
- Wade, W. A., Treat, T. A., & Stuart, G. L. (1998). Transporting an empirically supported treatment for panic disorder to a service clinic setting: A benchmarking strategy. *Journal of Consulting and Clinical Psychology, 66*, 231–239.
- Wampold, B. E. (1997). Methodological problems in identifying efficacious psychotherapies. *Psychotherapy Research, 7*, 21–43.
- Wampold, B. E. (2001). *The great psychotherapy debate: Model, methods, and findings*. Mahwah, NJ: Erlbaum.
- Weersing, V. R., & Weisz, J. R. (2002). Community clinic treatment of depressed youth: Benchmarking usual care against CBT clinical trials. *Journal of Consulting and Clinical Psychology, 70*, 299–310.
- Weissman, M. M., & Bothwell, S. (1976). Assessment of social adjustment by patient self-report. *Archives of General Psychiatry, 33*, 1111–1115.
- Weisz, J. R., Jensen-Doss, A., & Hawley, K. M. (2006). Evidence-based youth psychotherapies versus usual clinical care: A meta-analysis of direct comparisons. *American Psychologist, 61*, 671–689.
- Westen, D., & Morrison, K. (2001). A multidimensional meta-analysis of treatments for depression, panic, and generalized anxiety disorder: An empirical examination of the status of empirically supported therapies. *Journal of Consulting and Clinical Psychology, 69*, 875–899.
- Westen, D., Novotny, C. M., & Thompson-Brenner, H. (2004). The empirical status of empirically supported psychotherapies: Assumptions, findings, and reporting in controlled clinical trials. *Psychological Bulletin, 130*, 631–663.
- Wilson, S. B., Mason, T. W., & Ewing, M. J. M. (1997). Evaluating the impact of receiving university-based counseling services on student retention. *Journal of Counseling Psychology, 44*, 316–320.
- Wolgast, B. M., Rader, J., Roche, D., Thompson, C. P., von Zuben, F. C., & Goldberg, A. (2005). Investigation of clinically significant change by severity level in college counseling center clients. *Journal of College Counseling, 8*, 140–152.
- Zlotnick, C., Elkin, I., & Shea, M. T. (1998). Does the gender of a patient or the gender of a therapist affect the treatment of patients with major depression? *Journal of Consulting and Clinical Psychology, 66*, 655–659.

Received September 28, 2007

Revision received January 28, 2009

Accepted January 28, 2009 ■